Machine Learning-Based System for Document Management Optimization: A Case Study of a Small and Medium-Sized Enterprise

Oscar Mansilla*, Ronald Lopez-Haya

Universidad Tecnológica del Perú, Lima, Perú

Abstract

This study addresses inefficiencies in administrative document management at Kazoku No Mizu SAC (Peru), where manual transcription processes create operational bottlenecks, resulting in an average processing time of 70.64 seconds per document. Despite advances in Optical Character Recognition (OCR), conventional approaches exhibit limitations in handling noisy handwritten data, particularly in resource-constrained environments such as small and medium-sized enterprises (SMEs). To address this gap, the study proposes a hybrid Intelligent Document Processing (IDP) system based on Machine Learning that integrates computer vision through YOLOv8 for region-of-interest detection and a Large Language Model (Gemini 2.5 Flash API) for semantic validation and intelligent data recovery.The research adopts an applied, quantitative, non-experimental, and longitudinal design, analyzing a sample of 119 documents ($n = 119$). System performance was evaluated by comparing manual and automated processing outcomes using a paired-samples Student's t-test. The results demonstrate a statistically significant reduction in average processing time, from 70.00 seconds to 3.86 seconds per document ($t(118) = 90.21$, $p < 0.05$), achieving a temporal efficiency factor of 18.1 relative to the manual approach. The Automation Rate (AR) reached 98.66%, while the Intelligent Recovery Rate (IRR) achieved 89.77%, enabling the recovery of ambiguous information beyond the capabilities of conventional OCR systems. Furthermore, Data Accuracy (DA) reached 66.39%, representing a substantial improvement over the baseline scenario. However, inferential analysis revealed that human validation achieved significantly higher accuracy levels (89.83%) than the automated system (66.39%) ($t(118) = -10.45$, $p < 0.001$).The findings indicate that hybrid AI architectures significantly enhance operational efficiency while providing effective semantic support for document interpretation and recovery. The proposed solution aligns with the quality characteristics established in ISO/IEC 25010 and ISO/IEC 25012 standards, highlighting the potential of cost-effective IDP systems to optimize high-volume document processing workflows in SMEs operating under resource constraints.

Keywords

Machine Learning, Intelligent Document Processing, Computer Vision, Document Management, Process Optimization

1. Introduction

In the field of document management, organizations face significant challenges related to the efficient use of human, financial, and time resources, particularly in environments where processes rely on the manual or semi-automated handling and recording of information. Studies such as those conducted by Davenport and Kim [1] and Hammer and Stanton [2] indicate that inefficiencies in document management systems can lead to operational delays, increased costs, and errors in processed information, ultimately affecting service quality and organizational decision-making. From a quantitative perspective, reports published by McKinsey & Company [3] indicate that advances in automation and generative AI are expected to substantially increase the automation potential of administrative and information-processing activities, particularly those involving data collection, processing, and routine decision-making tasks. These findings suggest that organizations with highly document-intensive processes continue to allocate a considerable share of their operational effort to information-handling activities that could potentially be streamlined through intelligent automation technologies.

IWSAI 2026: 1st International Workshop on Systems, Applications, Artificial Intelligence and Innovation, September 16–18, 2026, Piura, Peru

*Corresponding author.

omansilla@utp.edu.pe (O. Mansilla)

0009-0003-2740-1517 (O. Mansilla); 0009-0007-5957-2475 (R. Lopez-Haya)

From an operational perspective, document-intensive processes often require considerable organizational effort due to the manual handling, validation, and registration of information. In this context, Abidemi [4] argues that small and medium-sized enterprises (SMEs) face additional challenges due to their limited technological resources and lack of process standardization, which increases their reliance on manual tasks and undermines their competitiveness. Furthermore, these constraints often hinder the adoption of advanced digital solutions, resulting in lower levels of operational efficiency, greater susceptibility to human error, and slower organizational responsiveness in document-intensive environments.

In response to this scenario, Abidemi [4] highlights that document automation based on Intelligent Document Processing (IDP) has emerged as a key solution for reducing operational processing times, minimizing errors, and improving information quality, enabling productivity gains of up to 30% in small and medium-sized enterprise (SME) environments.

In particular, processes such as document digitization, data capture, and validation constitute critical stages within the workflow, where a lack of standardization and variability in execution can generate operational bottlenecks. These challenges typically manifest as prolonged processing times, frequent rework, and the underutilization of available technological resources [5].

However, despite the availability of theoretical approaches aimed at process optimization and the development of document automation technologies, the integration of these two domains remains limited in specific organizational contexts, particularly within small and medium-sized enterprises (SMEs). In many cases, technological solutions are implemented without a structured methodology that supports process standardization and maximizes operational efficiency, thereby limiting their actual impact on organizational performance improvement.

In this regard, a knowledge gap is identified in the practical application of scientific management principles within digital document management environments, particularly in organizations whose administrative processes rely heavily on information handling and processing. Therefore, there is a need to develop solutions that integrate theoretical frameworks with technological tools tailored to the specific requirements of these organizations.

In this context, the scientific management approach proposed by Taylor [6] remains highly relevant, as it argues that organizational efficiency can be achieved through the application of scientific methods that enable the decomposition of complex tasks into elementary activities, the standardization of processes, and the optimization of resource utilization. However, within contemporary digital environments, these principles have evolved toward Business Process Management (BPM) and digital transformation paradigms, where organizational performance is enhanced through the integration of process standardization, data-driven decision-making, and intelligent technologies. Recent research highlights that BPM has become a strategic enabler of digital transformation, with artificial intelligence increasingly influencing the redesign, automation, and continuous improvement of business processes [7]. Accordingly, technologies such as Intelligent Document Processing (IDP) integrate Optical Character Recognition (OCR), Machine Learning (ML), and Natural Language Processing (NLP) to automate the capture, classification, validation, and analysis of document-based information. In this regard, IDP not only reduces manual intervention but also improves process accuracy, speed, and consistency, enabling the optimization of document workflows in organizations that depend heavily on unstructured information [8]. Therefore, the classical efficiency principles proposed by Taylor [6] can be reinterpreted through a contemporary BPM and artificial intelligence-driven perspective, where operational excellence is achieved through the systematic alignment of technology, organizational processes, and human resources. From this standpoint, document process optimization requires not only the adoption of advanced technological solutions but also the implementation of structured management frameworks that establish standardized procedures, minimize reliance on individual judgment, and support continuous process improvement. Consequently, efficiency is achieved through the integration of intelligent automation, process governance, and quality-oriented management practices.

In the case of Kazoku No Mizu SAC, a company engaged in water supply services, a significant bottleneck has been identified within its administrative operations, particularly in the document transcription and registration processes required to support its core operational activities. This situation

results in delays in customer service, an increased workload for administrative personnel, and a higher risk of inconsistencies in recorded information. Furthermore, a lack of standardized procedures and limited utilization of available technological resources have been observed, directly affecting the organization's operational efficiency.

Consequently, it is necessary to analyze and optimize the workflow associated with these processes through the application of scientific management principles and technological tools aimed at document automation. In this regard, the present study seeks to implement an automation solution based on Machine Learning techniques to optimize both the efficiency and effectiveness of the Document Registration Process (DRP) within the organization under study. Furthermore, the study aims to evaluate the levels of process efficiency and effectiveness before and after the proposed implementation, as well as to design and deploy an automation model that supports the continuous improvement of administrative operations.

2. Related Work

Recent literature demonstrates significant advances in document management automation through hybrid architectures that integrate Optical Character Recognition (OCR), Computer Vision, and Language Models. In this context, several studies have addressed the optimization of administrative processes through the digitization of physical documents.

They demonstrated that the integration of OCR with automation platforms enhances operational traceability and reduces delays in logistics environments [9]. However, their approach is limited to structured data and does not address semantic ambiguity in handwritten documents. Along the same line, Pingili [10] reported processing time reductions of up to 80% through the implementation of Intelligent Document Processing (IDP), demonstrating substantial improvements in operational efficiency. Nevertheless, these systems remain critically dependent on input quality, which constrains their performance in real-world scenarios.

With regard to Computer Vision, recent studies have demonstrated the growing effectiveness of multimodal document understanding architectures that combine visual, spatial, and textual features to improve information extraction from complex documents. Appalaraju et al. [11] and Bhattacharyya et al. [12] reported significant advances in document understanding through models capable of integrating visual and semantic representations. However, these approaches typically require substantial computational resources and large-scale training datasets, which may limit their applicability in small and medium-sized enterprises (SMEs) operating under resource-constrained conditions.

On the other hand, Ni et al. [13] and Rosli et al. [14] address the enhancement of document preprocessing through GAN-based models, achieving improvements in visual quality and facilitating subsequent information extraction. However, these approaches primarily focus on image enhancement and noise reduction rather than on the semantic recovery of data.

Regarding document understanding models, Appalaraju et al. [11] and Bhattacharyya et al. [12] introduced multimodal architectures capable of integrating visual, spatial, and textual information, achieving improvements across standard evaluation metrics. Nevertheless, these models exhibit high computational complexity and limited applicability within small and medium-sized enterprises (SMEs). Furthermore, Memon et al. [15] demonstrated that, despite advances in deep learning-based OCR, accuracy declines when processing continuous handwritten text, highlighting the need for contextual validation mechanisms. In this regard, Mubeen et al. [16] reported that tools such as Tesseract deliver satisfactory performance under controlled conditions but experience performance degradation when faced with document variability. Additionally, recent studies have shown that Vision Transformers are capable of capturing complex patterns in documents, achieving accuracy rates above 98%. However, these approaches are generally designed for specific tasks, such as document authentication, and do not address end-to-end document management processes [17].

At the national level, Osorio and Ruiz [18], Rafael et al. [19], and Ochoa and Osorio [20] demonstrate the feasibility of applying Optical Character Recognition (OCR) and Robotic Process Automation (RPA)

within organizational environments, achieving improvements in operational efficiency and reductions in error rates. However, these approaches primarily focus on process-level automation and do not incorporate advanced semantic inference mechanisms capable of enhancing the interpretation and validation of extracted information.

In summary, the literature reveals three major limitations: (i) dependence on input quality, (ii) limited semantic interpretation capabilities, and (iii) low adaptability in resource-constrained environments. In response to these challenges, the present study proposes a hybrid architecture that incorporates a semantic validation layer based on Large Language Models (LLMs), specifically designed for real-world environments characterized by high document variability.

3. Research Methodology

3.1. Case Study

Kazoku No Mizu S.A.C., located in Pisco, Peru, is a company dedicated to supplying water for logistics operations, primarily at the Port of Paracas. The company has provided explicit authorization for the use of its real corporate name, exact location, and operational data for academic and scientific research purposes, as well as for the publication of information related to the present study. Its operational structure combines field activities, including the reception and recording of tally sheets at the port, with administrative activities focused on office-based data consolidation.

The core business process involves the manual transcription of "tally sheet" forms, which are physical documents used to record critical variables such as vehicle license plate numbers, operating schedules, and cargo weights.

Prior to the intervention, the workflow began with the manual completion of these documents by port operators, followed by their physical transfer and subsequent data entry into Excel spreadsheets. This duplicated process resulted in an average processing time of 70.64 seconds per record, in addition to errors caused by poor handwriting, low ink quality, and missing information. It was determined that 32.39% of the documents exhibited legibility issues, adversely affecting data traceability and reliability.

Under these conditions, the company operated with a high dependence on manual labor, an absence of automation mechanisms, and limited responsiveness, constituting a representative example of the operational constraints commonly faced by small and medium-sized enterprises (SMEs). The study was conducted using organizational information provided and authorized by the company, ensuring compliance with ethical considerations regarding data access, disclosure, and publication.

3.2. Research Design

The study employs a deductive method, as it begins with general principles related to process optimization and automation and subsequently applies them to a specific organizational context through the analysis of the dimensions associated with the study variables. The research design is non-experimental because the variables are not deliberately manipulated but are instead observed in their natural setting before and after the technological implementation. Furthermore, the study follows a longitudinal approach, considering that data were collected at two distinct points in time: a pre-test phase (manual execution) and a post-test phase (automated execution). To reduce potential biases associated with learning effects or personnel adaptation, the pre-test measurements were obtained from the company's standard operational procedure before the implementation of the automated solution, while the post-test measurements were collected only after the system had been fully deployed and stabilized in routine operation. Additionally, the same document processing tasks, evaluation criteria, and performance indicators were used in both phases, ensuring comparability between measurements and minimizing the influence of external factors unrelated to the implemented solution. The scope of the research is descriptive-comparative, as it aims to describe and compare the characteristics of the process under both conditions.

3.3. Population and Sample

The study population consisted of 176 "tally sheet" forms generated by the company between November 2025 and January 2026, associated with water supply and maintenance operations. These documents constituted the unit of analysis and exhibited varying levels of legibility, enabling the evaluation of the proposed system under real-world operational conditions.

To determine the sample size, a census sampling approach was employed, initially considering the entire population ($N = 176$). However, based on the established exclusion criteria, documents containing overlapping handwriting or lacking visible ink were removed from the dataset. As a result, 57 documents were excluded, yielding a final sample of $N = 119$. This procedure was adopted to minimize methodological bias associated with partially or completely illegible documents.

3.4. Data Collection

Regarding data collection, three primary techniques were employed. First, a retrospective interview was conducted with personnel from the commercial department to estimate the initial operational workload, identifying an average manual processing time of 70.64 seconds per record. Second, a document analysis was performed on the collected tally sheet forms, which served as the primary data source for training the computational models and validating their outputs. Third, automated observation was carried out through system-generated logs to capture performance metrics such as execution time, resource consumption, and model accuracy.

The instruments used included an interview guide, a time-tracking sheet for comparing manual and automated performance, and a semantic validation matrix integrated into the system. This matrix contrasts the outputs generated by Optical Character Recognition (OCR) with the corrections provided by a Large Language Model (LLM), enabling the assessment of data accuracy and reliability.

To establish the ground truth used for evaluating Data Accuracy (DA), each tally sheet included in the sample was manually reviewed and transcribed by the researchers using the original physical documents as the reference source. The validation process consisted of comparing all extracted fields against a manually verified dataset created prior to the automated processing phase. In cases involving unclear handwriting, low-visibility ink, or ambiguous characters, the researchers cross-checked the information with related fields contained within the same document and with the company's operational records when available. The ground truth dataset was finalized before the execution of the automated system, ensuring that the reference values were independent from the outputs generated by the OCR, heuristic rules, or Large Language Model components. This procedure minimized confirmation bias and prevented circular validation during system performance assessment.

The retrospective interview followed a semi-structured format and was administered to personnel responsible for the document management process, with the objective of identifying the operational conditions that existed prior to the intervention. The interview questions were aligned with the efficiency and effectiveness dimensions of the study, enabling the collection of information regarding processing times, workload volume, error frequency, and data quality. This information was used to establish the baseline conditions of the process and to contextualize the results obtained following the implementation of the proposed solution, as shown in Table 1.

The document analysis consisted of a systematic review of physical tally sheets and digitized records, considering variables such as field completeness, information consistency, and document legibility. This instrument enabled the quantification of key indicators, including Data Accuracy (DA) and the Automation Rate (AR), thereby establishing an objective basis for comparing system performance before and after the intervention.

The observation technique was employed to directly monitor the operational workflow of the transcription process, identifying execution times and potential bottlenecks. Through this technique, Operational Time (Et) was measured for both the manual and automated approaches, enabling an assessment of the impact of the implemented solution on process efficiency.

Table 1
Alignment of Interview Questions with the Associated Dimensions and Indicators.

Evaluated Dimension	Associated Indicator	Questions
Operational Efficiency	Processing Time (Et)	• How long does it take to register a tally sheet? • Are there any delays in the registration process?
Operational Efficiency	Operational Volume	• How many tally sheets are processed per work shift? • Is there any accumulation of pending documents?
Process Effectiveness	Error Frequency	• How frequently do data entry errors occur? • Are data omissions observed during the registration process?
Process Effectiveness	Legibility	• Do the documents present legibility issues? • What are the most frequently encountered problems?

3.5. Study Variables and Dimensions

Process efficiency is defined, according to the International Organization for Standardization [21], as the relationship between the results achieved and the resources utilized, considering time as the critical resource in document processing activities. In this study, efficiency is assessed based on the system's ability to reduce human intervention and optimize processing time without compromising output quality. Two indicators are employed to measure this construct.

The first indicator is the Automation Rate (AR), which measures the percentage of fields automatically completed by the system:

Equation (1):

$$AR = \frac{Fields\ automatically\ completed\ by\ the\ algorithm}{Total\ required\ fields} \tag{1}$$

The second indicator is Relative Time Efficiency (Et), which compares the average manual processing time with the average automated processing time:

Equation (2):

$$ET = \frac{Average\ Manual\ Processing\ Time}{Average\ Automated\ Processing\ Time} \tag{2}$$

On the other hand, effectiveness is defined as the degree to which the intended outcomes are achieved, with a particular emphasis on the quality and validity of the generated information. In this context, effectiveness is evaluated based on the semantic accuracy of the extracted data. Two indicators are used to assess this construct.

The first indicator is the Intelligent Recovery Rate (IRR), which measures the system's ability to correct or infer illegible data:

Equation (3):

$$IRR = \frac{(AI\ Corrected\ Data + Data\ recovered\ through\ heuristic\ rules)}{Total\ processed\ data} \tag{3}$$

The second indicator is Data Accuracy (DA), which evaluates the degree of correspondence between the processed data and the validated ground truth:

Equation (4):

$$DA = \frac{Correctly\ Classified\ Fields}{Total\ processed\ data} \tag{4}$$

Together, these indicators provide a comprehensive assessment of the quality and reliability of the data generated by the system, ensuring alignment with internationally recognized software and data quality standards.

YOLOv8 was selected as the object detection architecture due to its favorable balance between detection speed, computational efficiency, and deployment feasibility in resource-constrained environments such as small and medium-sized enterprises (SMEs). Although alternative architectures and advanced data augmentation strategies may achieve higher detection accuracy under larger datasets, the objective of the present study was not to maximize object detection performance in isolation but rather to evaluate the effectiveness of a complete Intelligent Document Processing pipeline under real operational constraints. Preliminary analysis revealed that the primary limitation was not the localization of regions of interest, but the semantic ambiguity and variability of handwritten entries. Therefore, YOLOv8 was retained as a lightweight localization component, while semantic recovery and validation were delegated to the heuristic and Large Language Model layers. This design decision prioritizes practical applicability, computational affordability, and scalability over maximizing mAP performance alone.

3.6. Research Hypotheses

To validate the results, statistical hypotheses were formulated to evaluate the impact of the automated system on the levels of efficiency and effectiveness. Regarding efficiency, the null hypothesis (H_{01}) states that there is no significant increase in process efficiency following the implementation of the automated system, whereas the alternative hypothesis (H_{11}) states that there is a significant increase in process efficiency. Regarding effectiveness, the null hypothesis (H_{02}) states that there is no significant improvement in process effectiveness, whereas the alternative hypothesis (H_{12}) states that there is a significant improvement in process effectiveness.

Since the measurements were obtained from the same sample at two different points in time (pre-test and post-test), the Kolmogorov–Smirnov test ($n = 119$) was first applied to assess data normality. Subsequently, if the differences between paired observations followed a normal distribution, a paired-samples Student's t-test was employed; otherwise, the Wilcoxon signed-rank test was used. A significance level of $\alpha = 0.05$ was adopted to determine whether the observed differences were statistically significant and attributable to the implementation of the automated system.

Finally, data analysis was conducted using both descriptive and inferential statistics. Descriptive statistical measures, including means, standard deviations, coefficients of variation, and percentages, were used to summarize the data. Inferential statistical techniques were then applied to test the proposed hypotheses and validate the impact of the technological solution on the evaluated process.

4. Results

This section presents the results according to the structure of the proposed specific objectives. First, the findings associated with Specific Objective 1 are analyzed, focusing on process efficiency through the evaluation of automation performance and operational time reduction. Subsequently, the results related to Specific Objective 2 are presented, addressing process effectiveness in terms of intelligent data recovery and data accuracy. Finally, the findings are integrated to address the overall objective of the study.

4.1. Results for Specific Objective 1: Process Efficiency (DRP)

With regard to the objective of determining the level of efficiency before and after the implementation of the automated system, the results are first presented in terms of the Automation Rate (AR) and subsequently in terms of Relative Time Efficiency (Et), as discussed below.

4.2. Automation Rate (AR)

After processing the validated dataset, the system achieved an Automation Rate (AR) of 98.66%, automatically completing 1,174 out of 1,190 fields. This result represents a substantial improvement compared with the initial system versions, in which automation levels did not exceed 30%. Moreover, the automated system outperformed human performance, which achieved a Field Completion Rate (FCR) of 93.95% (see Figure 1). The system also exhibited a lower coefficient of variation, reaching 5.59% compared with 7.23% for the human-operated process.

These findings indicate that the automated system delivers greater operational consistency. However, it remains sensitive to extreme cases, particularly in Documents 10, 25, and 115. In contrast, the human-operated process demonstrates greater flexibility but lower consistency, exhibiting higher variability while experiencing less pronounced performance declines in atypical cases.

This high level of functionality confirms that the implemented architecture effectively minimizes manual intervention in document registration, in accordance with the functional suitability criteria established in the ISO/IEC 25010 standard [22].

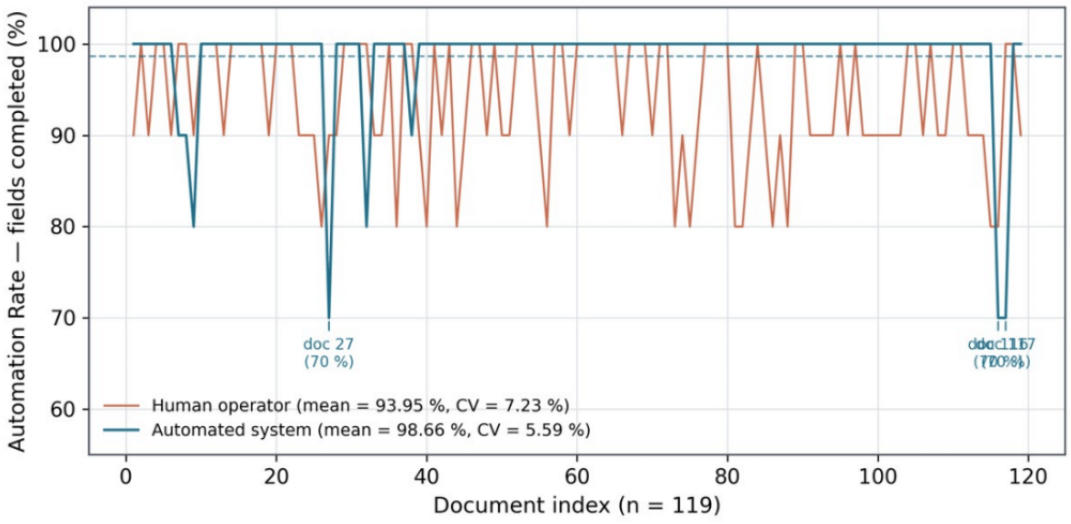

Figure 1: Comparison of the Automation Rate (AR) Between the Automated System and Human Performance.

4.3. Relative Time Efficiency (Et)

The time analysis revealed a significant improvement in processing speed. While the manual approach required an average of 70.64 seconds per ticket, the automated system reduced the average processing time to 3.86 seconds per ticket (Figure 2).

As a result, the Relative Time Efficiency (Et) reached a value of 18.1, indicating that the automated system processed documents approximately 18 times faster than the manual approach, as shown in the following calculation:

$$ET = \frac{70.0s}{3.86s} \tag{5}$$

This value indicates that the system is capable of processing approximately 18 documents in the time previously required to process a single document manually, thereby eliminating the operational bottleneck identified in the baseline process.

Furthermore, the progression from a linear-processing version (26.95 seconds per document) to the optimized “Turbo” version highlights the effectiveness of implementing batch-processing and multithreading techniques in resource-constrained environments.

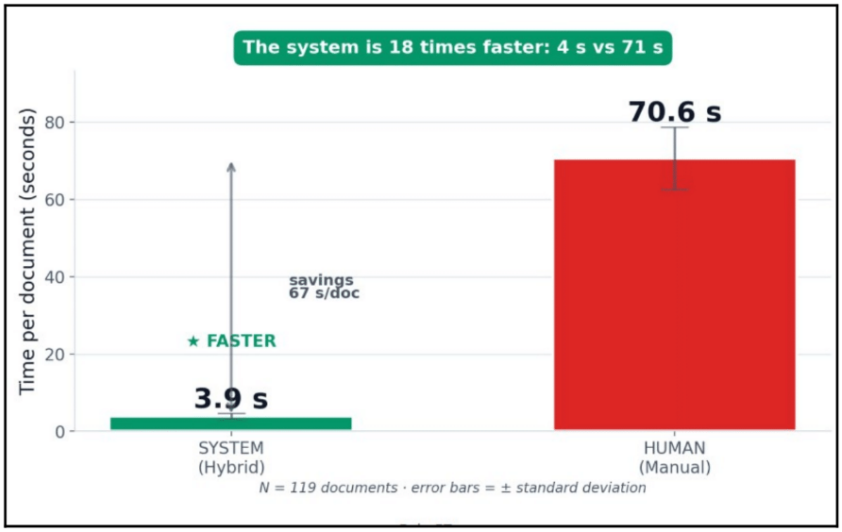

Figure 2: Operational Efficiency Analysis Before and After Automation.

For the Relative Time Efficiency analysis, a statistical assessment was conducted to determine whether significant differences existed between the document processing times achieved by the automated system and those obtained through manual processing. To this end, processing times from 119 events (document-reading tasks) were collected. Given the paired nature of the observations, the difference between the human processing time and the machine processing time was calculated for each event. Subsequently, the Kolmogorov–Smirnov test was applied to the resulting 119 paired differences to assess data normality. The test yielded a K–S statistic of 0.0739 with a p-value of 0.5101 at a significance level of 0.05, indicating that the distribution of differences satisfied the normality assumption.

To determine whether a statistically significant difference existed between the two approaches, a paired-samples Student's t-test was then performed to compare processing times under the human-operated and automated methods. The results revealed statistically significant differences between the two groups, with a test statistic of $t(118) = 90.21$ and $p < 0.05$. Specifically, the average processing time of the manual method ($mean = 70.64$; $SD = 8.11$) was substantially higher than that of the automated method ($mean = 3.86$; $SD = 0.81$), demonstrating that the proposed technological solution significantly improved process efficiency.

Overall, the results confirmed that the implementation of the hybrid Machine Learning-based system significantly enhanced the operational efficiency of the document management process, as evidenced by a substantial increase in the Automation Rate (AR) and a dramatic reduction in processing time. The statistically significant difference observed between the manual and automated approaches validates the effectiveness of the proposed intervention, establishing the system as a robust solution for eliminating operational bottlenecks in real-world environments.

4.4. Results for Specific Objective 2: Process Effectiveness (DRP)

With regard to the evaluation of process effectiveness, the indicators employed were the Intelligent Recovery Rate (IRR) and Data Accuracy (DA), as detailed in the following sections.

4.5. Intelligent Recovery Rate (IRR)

The system achieved an Intelligent Recovery Rate (IRR) of 89.77%, demonstrating its ability to recover and correct most records that would otherwise be classified as ambiguous or erroneous by conventional OCR systems. This result highlights the value of integrating Machine Learning techniques with semantic validation mechanisms, allowing the system to operate effectively under adverse conditions such as poor handwriting, visual noise, and low ink quality. Regarding the Intelligent Recovery Rate (IRR),

this indicator measures the capability of the hybrid system operating in its "Turbo" mode to restore logistics-related information through a recursive learning process. In this architecture, YOLOv8 is used exclusively for the spatial localization of document regions, namely Regions of Interest (ROIs), while information recovery relies on higher-level inference mechanisms. As shown by the training curves (Figure 3), the limited dataset imposed a practical performance ceiling, reflected by a maximum mAP50 value of 0.25, preventing reliable extraction based solely on rigid bounding-box detection. This relatively low mAP50 value should be interpreted within the context of the study objective. The system was designed as a hybrid document-processing architecture rather than a standalone object-detection model. Accordingly, YOLOv8 functioned only as a region-localization mechanism, while information extraction depended primarily on the heuristic and semantic inference layers. To mitigate this limitation, the proposed solution incorporates a Python-based heuristic layer that performs indexed queries on an incremental JSON knowledge repository (learned_patterns.json). This component leverages previously acquired patterns and contextual information to recover incomplete, ambiguous, or poorly recognized fields, thereby improving information retrieval performance beyond the capabilities of conventional CR based approaches.

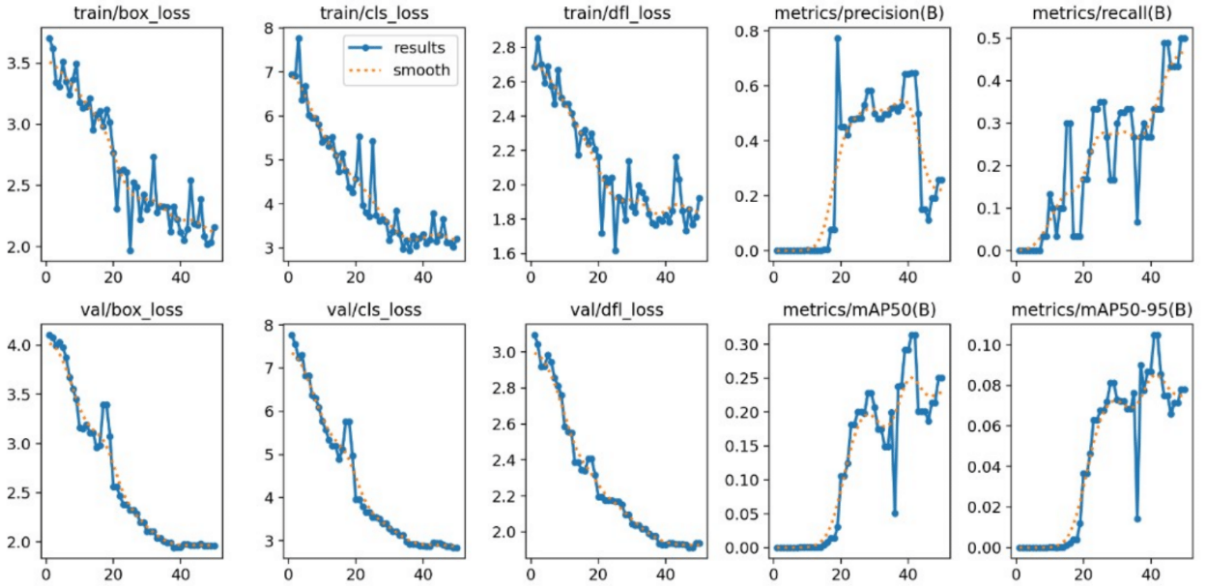

Figure 3: Training and Validation Curves and Performance Metrics of the YOLOv8 Model During the Learning Process.

Figure 4 presents the decomposition of the Intelligent Recovery Rate (IRR) into its heuristic, artificial intelligence, and integrated components. The results show that the heuristic rule-based subsystem recovered 21.59% of the data (26 out of 119 records), demonstrating a limited yet meaningful contribution to the correction of structured errors and predictable patterns.

In contrast, the artificial intelligence layer (Gemini) achieved a substantially higher recovery rate of 68.18% (81 out of 119 records), demonstrating its ability to resolve semantic ambiguities and reconstruct information from incomplete or degraded contexts.

The integration of both approaches resulted in a total IRR of 89.77% (107 out of 119 records), evidencing a substantial improvement in data recovery performance compared with the isolated use of either method.

These findings confirm the effectiveness of the proposed hybrid approach, in which heuristic rules function as a baseline correction mechanism, while the artificial intelligence component contributes advanced inference capabilities. Together, these components create a robust system capable of handling inputs characterized by high variability, ambiguity, and noise.

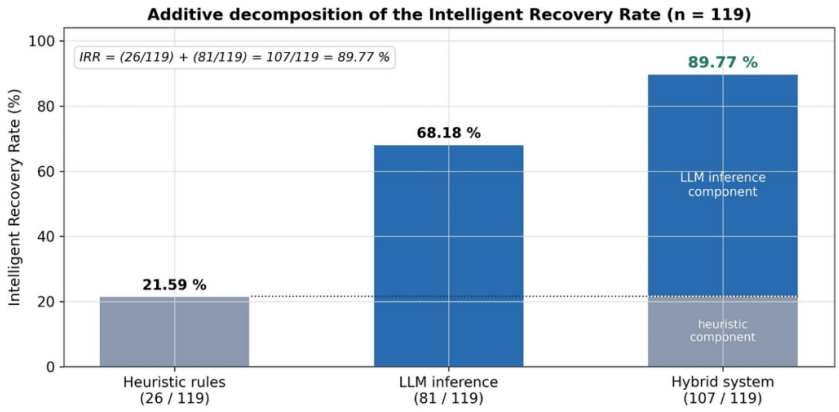

Figure 4: Intelligent Recovery Rate (IRR) Decomposed into Heuristic, Artificial Intelligence, and Hybrid Integration Components.

4.6. Data Accuracy (DA)

During the initial phase of the project, the system achieved a Data Accuracy (DA) of only 22.9%, primarily due to the absence of learned patterns and the high variability of the input data.

Following the implementation of an incremental learning model based on pattern storage (learned_patterns.json) and semantic validation through a Large Language Model (LLM), data accuracy increased to 66.39%, representing an improvement of 43.49 percentage points.

This improvement confirms the system's ability to learn progressively and optimize its performance in real-world operational environments, meeting the semantic accuracy requirements established by the ISO/IEC 25012 standard.

Regarding Data Accuracy (DA), the frequency distribution analysis conducted on the evaluated sample ($n = 119$ documents) revealed that the automated system outperformed the human operator in 13% of the cases and achieved parity in an additional 12% of the evaluated interactions. Although human performance remained superior in the remaining 76% of cases, a condition largely attributable to the high heterogeneity and handwriting variability present in the source documents, the evolution of the model relative to its initial baseline DA of 22.9% is particularly noteworthy.

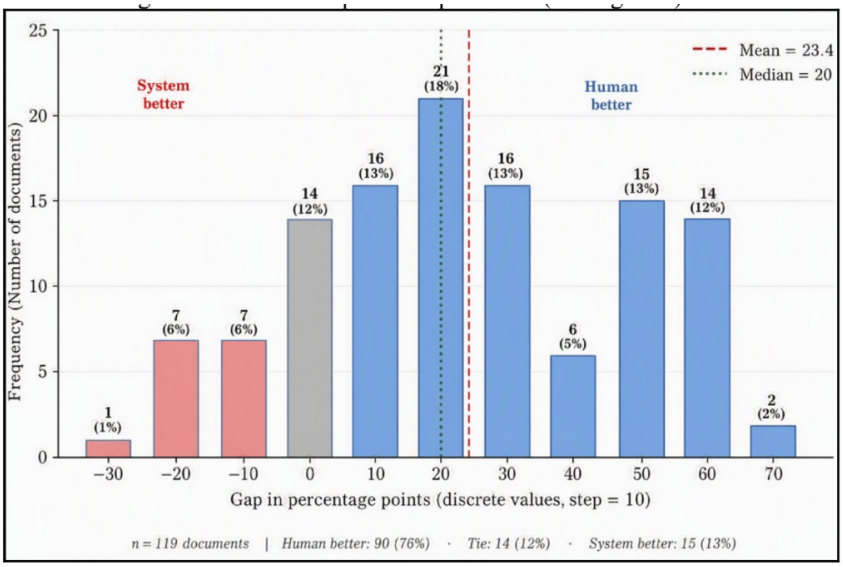

Figure 5: Distribution of the Performance Gap between the Automated System and Human Performance in Data Accuracy.

The observed increase of 43.49 percentage points demonstrates that the implemented recursive

learning mechanism effectively mitigates the limitations inherent to a small dataset, a scenario commonly encountered in small and medium-sized enterprises (SMEs). Consequently, the adoption of this architecture represents a strategic advancement, as it enables the reconfiguration of human roles from routine data-entry activities toward data validation and quality assurance tasks. This shift substantially reduces both operational processing time and the cognitive workload imposed on personnel (see Figure 5).

Figure 6 illustrates the accuracy achieved for each processed document by comparing the performance of the automated system with that of the human operator. As shown, the human-operated process exhibits high accuracy and stability, with a mean accuracy of 89.83% and a coefficient of variation of 9.61%. In contrast, the automated system demonstrates less consistent performance and a strong dependency on document characteristics, exhibiting frequent and pronounced performance declines that result in a mean accuracy of 66.39% and a coefficient of variation of 34.51%.

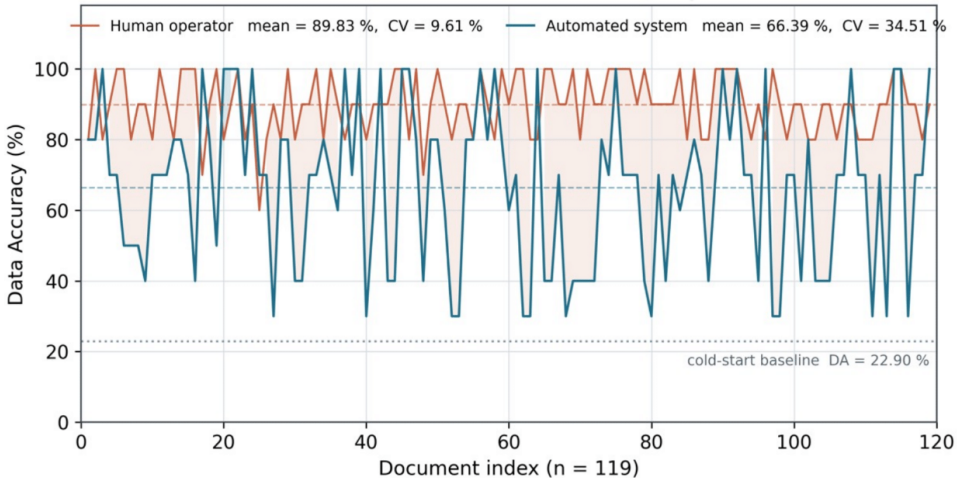

Figure 6: Comparison of Data Accuracy (DA) Between the Automated System and Human Performance.

This behavior may be primarily attributed to document legibility issues, considering that 32.39% of the evaluated documents were classified as partially or fully illegible. In addition, OCR technology operates predominantly at the visual recognition level without fully understanding semantic meaning, which limits its ability to resolve ambiguous or degraded content. Although this limitation could be mitigated through continuous autonomous learning, such learning was not fully achieved during the present pilot implementation due to the limited dataset available for meaningful model adaptation. Another contributing factor may be the lack of a standardized document structure, which introduces considerable variability in document completion and negatively affects extraction consistency.

Despite these limitations, the achieved Data Accuracy (DA) of 66.39% represents a substantial improvement over the initial baseline performance of 22.9%, demonstrating the effectiveness of the proposed hybrid architecture under real-world operational conditions characterized by document heterogeneity and variable data quality.

To determine the appropriate statistical test for hypothesis testing, the normality assumption was first assessed using the paired differences between the Data Accuracy (DA) scores obtained by the automated system and those achieved through human processing ($n = 119$). The Kolmogorov-Smirnov test yielded a statistic of $KS = 0.0739$ and a significance value of $p = 0.5101$. Since the p-value exceeded the established significance level $\alpha = 0.05$, the null hypothesis of normality was not rejected. Therefore, the paired differences were considered normally distributed, justifying the use of the paired-samples Student's t-test.

Subsequently, a paired-samples Student's t-test was performed to determine whether significant differences existed between the Data Accuracy (DA) achieved by the automated system and that obtained through human validation. The results showed a mean accuracy of 66.39% for the automated system and 89.83% for the human-operated process. The analysis produced a test statistic of $t(118) = -10.45$

with a two-tailed significance value of $p = 1.7110^{-18}$, indicating statistically significant differences between both methods. Since the p-value was lower than 0.05, the null hypothesis was rejected. The negative t-value indicates that the automated system achieved a significantly lower level of Data Accuracy than the human-operated process. Therefore, although the proposed solution substantially improved operational efficiency and automation performance, the results demonstrate that its data accuracy remains significantly below the level achieved through human validation under the evaluated conditions.

In summary, the results demonstrate that the proposed hybrid solution significantly improves process effectiveness by achieving a high Intelligent Recovery Rate (IRR), thereby confirming its ability to manage ambiguous data through semantic inference mechanisms. However, although a substantial improvement in Data Accuracy (DA) was achieved relative to the baseline condition, the system's performance remains below that of human operators, primarily due to the variability and quality of the input data.

Consequently, the proposed system positions itself as a strategic support tool that enhances information recovery and pre-processing activities, while human validation remains essential in scenarios characterized by high uncertainty. Overall, the findings highlight the potential of hybrid AI-driven architectures to augment human capabilities, improve data processing effectiveness, and support decision-making processes in document-intensive operational environments.

4.7. Results for the General Objective: Implementation of the Machine Learning Solution

The overall objective of the study was to implement a Machine Learning-based solution to optimize the effectiveness of human resources within the organization under study. In this regard, the obtained results demonstrated a significant improvement in the performance of the document registration process, particularly in terms of operational efficiency.

The implementation of the hybrid system, based on Computer Vision techniques, Optical Character Recognition (OCR), and Large Language Models (LLMs), enabled the automation of the tally sheet registration process, substantially reducing human intervention while improving the quality of the processed data. Furthermore, the solution successfully addressed the major challenges identified during the diagnostic phase, particularly those associated with processing latency and variability in document quality.

From an operational perspective, the system demonstrated the capacity to process up to 100 tickets per day under the constraints of a free-tier infrastructure, representing a considerable improvement over the traditional manual approach. This throughput was observed during real-world pilot testing and reflects the maximum number of documents processed without exceeding the API quotas and infrastructure limitations associated with the free-tier service.

The proposed solution was implemented using a hybrid architecture that integrates OCR, Computer Vision models (YOLOv8), and Large Language Models (Gemini 2.5 Flash), complemented by heuristic rules developed in Python.

One of the system's main technical contributions was the optimization of resource consumption through batch-processing strategies. The transition from a linear execution model to an optimized architecture reduced API requests by 79.3%, decreasing from 176 to 36 requests for the same dataset (Figure 7).

Additionally, total token consumption was reduced by 18%, ensuring the sustainability of the system within the limitations imposed by free-tier infrastructure. This characteristic makes the proposed solution particularly suitable for small and medium-sized enterprises (SMEs), where technological and financial resources are often constrained.

Overall, the findings demonstrate that the proposed hybrid Machine Learning architecture not only improves operational efficiency and process effectiveness but also provides a cost-effective and scalable alternative for document-intensive administrative environments. The integration of Computer Vision, OCR, heuristic reasoning, and Large Language Models enables substantial reductions in processing

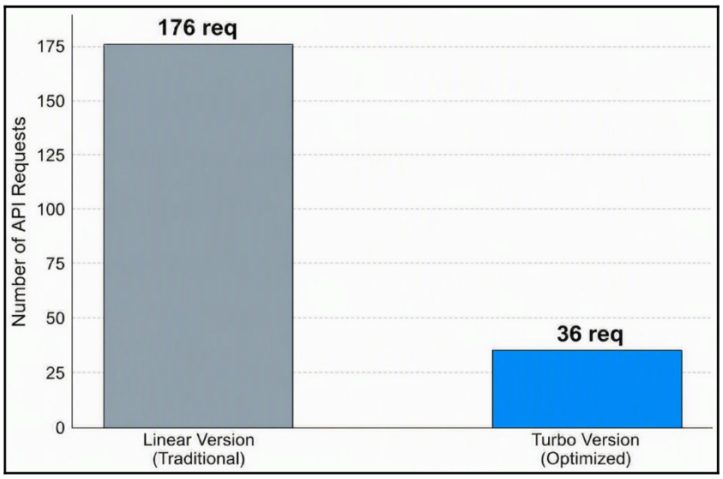

Figure 7: API Request Consumption in the Linear and Optimized Versions.

time, enhanced information recovery capabilities, and improved data quality, thereby supporting more efficient resource utilization and organizational decision-making.

In conclusion, the implementation of the hybrid Machine Learning system significantly enhanced the effectiveness of the document management process, primarily through improvements in operational efficiency and more rational resource utilization. The proposed architecture not only reduced processing times and computational costs but also demonstrated its viability in technologically constrained environments, establishing itself as a scalable solution applicable to small and medium-sized enterprises (SMEs) with similar administrative processes.

5. Discussion

The results demonstrated that the implementation of the hybrid Machine Learning system achieved substantial improvements in operational efficiency, although limitations in semantic accuracy remained when compared with human operators. Consistent with the findings of Pingili [10], Intelligent Document Processing (IDP) systems enable the automation of document workflows and significantly reduce processing times; however, their performance remains heavily dependent on input data quality and the semantic complexity of the content being processed.

In the present study, the 94.49% reduction in operational processing time and the high Automation Rate (AR) of 98.66% confirmed that the proposed solution substantially outperformed the manual approach in terms of efficiency. These findings are consistent with previous research on AI-based automation in document-intensive processes [9]. Nevertheless, the observed difference in Data Accuracy ($DA = 66.39\%$ versus 89.83% achieved by human operators) revealed a structural limitation in the processing of unstructured handwritten content. This result is consistent with the observations of Memon et al. [15], who emphasize that handwritten text recognition in real-world environments remains a challenging and unresolved problem, even for advanced deep learning models.

The high Intelligent Recovery Rate ($IRR = 89.77\%$) further validates the effectiveness of the proposed hybrid approach. While heuristic rules contributed a baseline recovery rate of 21.59%, the artificial intelligence component accounted for an additional 68.18%, demonstrating that contextual inference served as the primary recovery mechanism in ambiguous scenarios. This finding is consistent with Bhattacharyya el al. [12], who argue that semantic understanding based on language models outperforms traditional approaches focused primarily on document geometry. Furthermore, Appalaraju et al. [11] emphasize that the multimodal integration of visual and linguistic signals reinforces the need for hybrid architectures to enhance system robustness and overall information extraction performance.

Moreover, the approximately 79% reduction in API request consumption demonstrated the effectiveness of the optimized architecture based on batch-processing strategies. This outcome is consistent

with the performance efficiency principles established in the ISO/IEC 25010 quality model [22]. The result also supports the practical viability of the proposed system in resource-constrained environments, particularly within small and medium-sized enterprises (SMEs), as similarly reported by Osorio and Ruiz [18].

Overall, the findings confirmed that the proposed system does not completely replace human inferential capabilities but rather complements them. In environments characterized by high variability, ambiguity, and noise, the combination of algorithmic automation and human validation emerges as the most robust approach for ensuring both efficiency and quality in document management processes. Consequently, the proposed hybrid architecture should be viewed as an AI-augmented decision-support solution, in which automated processing enhances productivity while human expertise remains essential for handling complex and uncertain cases.

5.1. Study Limitations

Additionally, handwriting variability and the heterogeneous quality of the source documents constituted external and largely uncontrollable factors that significantly affected system accuracy, particularly in highly ambiguous scenarios. Furthermore, although YOLOv8 was incorporated as the region-of-interest detection component within the proposed architecture, the study did not document a formal training protocol including dataset partitioning strategies, hyperparameter optimization procedures, or transfer learning configurations. Consequently, it was not possible to perform a rigorous assessment of potential overfitting effects associated with the limited dataset size. However, it should be noted that YOLOv8 was not the primary object of investigation but rather a supporting localization mechanism within a broader hybrid Intelligent Document Processing architecture. Future research should incorporate fully documented training configurations, systematic benchmarking against alternative object detection models, and data augmentation strategies to improve reproducibility and facilitate a more comprehensive evaluation of model generalization performance.An additional limitation is related to the exclusion of 57 documents (32.39% of the initial population) that contained overlapping handwriting or lacked sufficient visual information for reliable processing and validation. Although this exclusion was necessary to ensure the availability of a verifiable reference dataset, it may have introduced a potential selection bias by removing the most challenging cases from the evaluation process. Consequently, the reported values for Data Accuracy (DA), Intelligent Recovery Rate (IRR), and other performance indicators should be interpreted as representative of documents that satisfy the minimum legibility conditions established in this study rather than of the entire universe of operational documents. Future research should investigate techniques specifically designed for severely degraded, incomplete, or highly ambiguous documents in order to assess system performance under more demanding real-world conditions. These limitations highlight the need for future improvements aimed at enhancing the robustness, transparency, adaptability, reproducibility, and generalization capabilities of the proposed model.

6. Conclusion

This research demonstrated that the implementation of a hybrid Machine Learning system significantly improves administrative document management, particularly in terms of operational efficiency. The reduction in processing time from 70.64 to 3.86 seconds per document (94.49%), together with the high Automation Rate (AR) of 98.66%, provides clear evidence of the effective elimination of bottlenecks associated with manual document registration. Furthermore, the observed improvement was statistically significant according to a paired-samples Student's t-test ($t(118) = 90.21$, $p < 0.001$), confirming the feasibility of deploying intelligent architectures in logistics environments characterized by high operational demand.

However, the results also showed that the effectiveness of the system, measured through Data Accuracy ($DA = 66.39\%$), remains lower than that of human operators (89.83%). This finding reveals a structural limitation in the processing of unstructured and handwritten data. Nevertheless, this gap does not invalidate the proposed solution; rather, it confirms that automation alone cannot guarantee semantic

accuracy, particularly when input data quality is poor. In this context, the high Intelligent Recovery Rate ($IRR = 89.77\%$) validates the contribution of the hybrid approach, in which artificial intelligence complements the limitations of conventional OCR through contextual inference mechanisms.

Overall, the findings confirm that the proposed model does not replace human intervention but rather transforms its role, positioning the system as a large-scale automation mechanism and the human operator as a validator of critical exceptions. This approach enables a balance between efficiency and reliability in real-world operational environments.

For future research, it is recommended to substantially expand the training dataset, incorporate more robust multimodal models, and explore advanced fine-tuning techniques specifically designed for handwritten text recognition. Furthermore, investigating the integration of continuous learning systems capable of progressively improving model accuracy according to operational context represents a promising research direction. Finally, further studies should focus on the design of standardized data-capture formats, as input data quality remains a critical factor influencing the overall performance of automated document processing systems.

Declaration on Generative AI

During the preparation of this manuscript, the authors used Microsoft Copilot to assist with language refinement, grammar checking, paraphrasing, and text improvement. After using this tool, the authors reviewed, edited, and validated the content as necessary and take full responsibility for the final version of the manuscript.

References

[1] T. H. Davenport, J. Kim, Keeping up with the quants: Your guide to understanding and using analytics, Harvard Business Review Press, 2013.
[2] S. A. Stanton, M. Hammer, The reengineering revolution: the handbook, Nicholas Brealey, 1995.
[3] McKinsey & Company, How automation is shaping the future of work, 2023. URL: https://www.mckinsey.com/featured-insights/themes/how-automation-is-shaping-the-future-of-work.
[4] A. Abidemi, The role of technology and automation in streamlining business processes and productivity for smes, International Journal of Entrepreneurship 7 (2024) 25–42.
[5] K. C. Laudon, J. P. Laudon, Management information systems: Managing the digital firm, New Jersey 8 (2004).
[6] F. W. Taylor, The principles of scientific management, NuVision Publications, LLC, 1911.
[7] B. Q. Truong, A. Nguyen-Duc, N. T. C. Van, A quantitative review of the research on business process management in digital transformation: a bibliometric approach, Software 2 (2023) 377–399.
[8] K. Kapula, Intelligent document processing: The new frontier of automation, International Journal of Intelligent Systems and Applications in Engineering 10 (2022) 215–223.
[9] J. Chuquival, R. Torres, M. Salazar, Automatización de procesos documentales mediante OCR y plataformas de integración en la gestión de combustible, Revista Internacional de Ingeniería Industrial 14 (2025) 23–34.
[10] P. Pingili, Intelligent document processing using machine learning and natural language processing, Journal of Information Systems Engineering 11 (2025) 45–58.
[11] S. Appalaraju, P. Tang, Q. Dong, N. Sankaran, Y. Zhou, R. Manmatha, Docformerv2: Local features for document understanding, in: Proceedings of the AAAI conference on artificial intelligence, volume 38, 2024, pp. 709–718.
[12] A. Bhattacharyya, A. Tripathi, U. Das, A. Karmakar, A. Pathak, M. Gupta, Information extraction from visually rich documents using llm-based organization of documents into independent textual segments, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 17241–17256.

[13] M. Ni, Z. Liang, J. Xu, Removal of color-document image show-through based on self-supervised learning, Applied Sciences 14 (2024) 4568.
[14] L. H. Rosli, Y. K. Hooi, K. B. Ong, Unsupervised document binarization of engineering drawings via multi noise cyclegan, International Journal of Advanced Computer Science and Applications 14 (2023).
[15] J. Memon, M. Sami, R. A. Khan, M. Uddin, Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr), IEEE access 8 (2020) 142642–142668.
[16] M. Mubeen, S. Ali, A. Khan, Implementation of Tesseract OCR for mobile-based text recognition systems, International Journal of Computer Applications 183 (2022) 15–22.
[17] Y. Hao, Z. Zheng, Research on detecting ai-generated forged handwritten signatures via data-efficient image transformers, IEEE Access 13 (2025) 7683–7690.
[18] S. G. Osorio Cruz, A. A. Ruiz Silva, Propuesta de automatización del proceso de validación y registro de facturas de servicios logísticos en operaciones de importación en una empresa siderúrgica., 2025.
[19] J. Rafael, P. Gómez, L. Torres, Sistema portátil basado en visión artificial para la lectura de documentos impresos, 2025.
[20] A. D. Ochoa Surco, P. E. Osorio Schuler, Implementación de un robotics process automation (rpa) para mejorar el proceso de validación de estados de cuenta en una entidad financiera, lima-2022, 2022.
[21] International Organization for Standardization, Iso 9000:2015 quality management systems – fundamentals and vocabulary, 2015.
[22] International Organization for Standardization, Iso/iec 25010:2011 systems and software engineering, 2011.